EFFECTIVE SENTIMENT ANALYSIS OF CORPORATE FINANCIAL REPORTS

Research-in-Progress

Jimmy SJ. Ren
Department of Information Systems
83 Tat Chee Ave, Hong Kong SAR
jimmy.sj.ren@gmail.com

Huizhong Ge
Department of Economics and Finance
83 Tat Chee Ave, Hong Kong SAR
lexi.ghz@gmail.com

Xiaoyu Wu
Department of Information Systems
83 Tat Chee Ave, Hong Kong SAR
xiaoyuwu5@gmail.com

Guan Wang
Department of Information Systems
83 Tat Chee Ave, Hong Kong SAR
guanwang3-c@my.cityu.edu.hk

Wei Wang
Department of Information Systems
83 Tat Chee Ave, Hong Kong SAR
wewang8@cityu.edu.hk

Stephen Shaoyi Liao
Department of Information Systems
83 Tat Chee Ave, Hong Kong SAR
issliao@cityu.edu.hk

Abstract

Sentiment analysis is widely adopted in studying various important topics in business intelligence. Though many studies reported interesting results by using machine learning, the lack of theoretic analysis and the shortage of practical guidance are hurdles of theory development. Besides, due to the difficulty in labelling data, the effectiveness of sentiment analysis with only labelled data needs to be questioned. In this paper, we drew on statistical learning theory to perform extensive theoretic analysis in sentiment analysis by using real corporate financial reports. We investigated when and why machine learning methods provide preferred performance under the guidance of the theory. We also provided practical suggestions in applying machine learning methods for both researchers and practitioners. In addition, we utilized the cheap and ubiquitous unlabelled data to further improve the sentiment analysis performance. This has the potential to largely reduce the manual data labelling work and to scale up the experiments.

Keywords: Text classification, sentiment analysis, machine learning, unlabeled data

Introduction

In recent years, many researchers in accounting, finance and information systems focused on coping with various kinds of business intelligence (BI) related research questions by analyzing companies' textual materials. We enumerate a few which have caught much attention. The list includes identifying companies' risk factors (Huang and Li 2011), predicting future earnings (Li 2010), predicting stock prices (Kravet and Muslu 2012) and detecting business frauds (Humpherys, et al. 2011), etc. From these studies, it's not hard to discern two facts and one trend. The first fact is that due to its openness and integrity, Form 10-K (required by U.S. Securities and Exchange Commission (SEC)) is widely used in addressing the aforementioned research questions. The second fact is that text based sentiment analysis, one of the most popular research approaches among these studies, is becoming a fundamental work in financial content analysis. One typical way of carrying out sentiment analysis is to classify every sentence of the Management's Discussion and Analysis (MD&A) in Form 10-K into positive, negative or neutral tone and then generate a quantitative measure for various tasks such as future earning prediction and stock price prediction. Meanwhile, there is also an emerging methodological trend in the literature, namely more researchers tend to adopt advanced machine learning (ML) methods in analyzing textual data and have already generated some interesting results.

We believe this phenomenon is a product of an important trend. With the rapid development of the Internet, we are now entering the era of "Big Data". As a result, the amount of business data (e.g. companies' textual materials) researchers exposed to is dramatically expanding. While there are many exciting knowledge discovery opportunities within the data, there is a call for more sophisticated and intelligent data analysis methods in uncovering complex patterns and structures of the data. As it is likely that the amount of available business data would continue to grow, it's reasonable to anticipate that intelligent methods such as machine learning methods would be more prevalent in the future studies.

However, despite we can find a number of studies adopted machine learning methods in this area, it's argued that applying such methods to study BI related topics in finance or accounting is still in its infancy (Huang and Li 2011). The main reasons, in our opinion, are the following. First of all, there is no previous study to our knowledge carried out an in-depth theoretic analysis on the methodology and provided a systematic explanation on when and why a particular machine learning method would work well. Secondly, all the previous studies in this field only took labeled data into account in the model training phase. This largely overlooks the potential power of unlabeled data which is not only plentiful but very easy to obtain as well. Thirdly, there is very few previous study compared different learning models and provided best practices in applying machine learning methods in different scenarios. We argue that without filling these three research gaps, it is not only easier for business researchers to be drowned by unobvious technical traps; it's also hard to promote a cumulative manner in the literature and therefore would be a hurdle in theory development (Dubin 1978).

The main goal of this paper is to address the aforementioned research gaps. We collected textual data from companies in very different industries. We conducted experiments using widely adopted text mining algorithms and showed different models would have their own merits in different situations. We also showed that by carefully adding unlabeled data, the classification performance is likely to be boosted. We not only provide theoretic analysis on when and why a particular model as well as adding unlabeled data would work well, we also provide practical guide in applying text mining algorithms in practice. To our knowledge, this paper is the pioneer in accumulating financial text mining knowledge and theoretical insights in a fundamental perspective in this field.

The structure of this paper is as follows. Section two reviews the extant literature and defines the research gaps. Section three presents the theoretical background, data collection process and text mining methodology, including how to utilize unlabeled data. Section four presents the experiment results and provides theoretic analysis to explain why the results are reasonable. Section five presents the practical suggestions in applying text mining methods then elaborates the research contributions and limitations. We conclude the paper and indicate the future work in section six.

Literature Review and Research Gaps

There is a considerable body of recent literature in accounting, finance and information systems, under the name of content analysis, addresses various corporate business problems by taking advantage of textual data from the public corporate annual reports, public news or social network. For example, Huang and Li (2011) developed a multi-label text classification algorithm to identify companies' risk factors by using textual information in SEC Form 10-K. Li (2010) examined the forward-looking statements in the MD&A section of Form 10-K and found that the average tone of the forward-looking statements is positively related to the company's future earnings. Kravet and Muslu (2012) also scrutinized the 10-K filings and found that the change in companies' textual risk disclosures is associated with the change in stock market. In Humpherys, et al. (2011), the authors investigated the deceptive language in the MD&A section of Form 10-K filings and found fraudulent disclosures tend to use more activation words but contain less lexical diversity. Tetlock et al. (2008) analyzed companies' IPO prospectuses and found simple quantitative measure of language has the predictive power of firms' accounting earnings and stock returns. Oh and Sheng (2011) investigated the content of micro blogs and found it has the predictive power over future stock price.

It's quite noticeable that it was not uncommon for researchers to adopt textual data from the 10-K filings in their studies. This is not only because Form 10-K is a required public corporate filing with SEC and provides comprehensive information of the companies' business conditions, it's also because the unaudited MD&A section in this document is the most read section and may contain rich insights and many potential research opportunities (Humpherys, et al. 2011).

However, it may not be obvious that within these studies, sentiment analysis (or tone classifications) seems to dominate the existing research approaches. Researchers use different methodologies to classify financial disclosure statements into several sentimental tones (Li 2010, Kothari et al. 2009, Bryan 1997, Davis et al. 2012), e.g. positive, negative or neutral. The output of the sentiment analysis well facilitates the business researchers to explore the financial or economic implications of the companies' filings. Such implication would be used as an important vehicle in addressing many important research questions including the aforementioned ones. Interested readers may refer to the comprehensive summary in (Li 2010) to know more details on this contention. In sum, sentiment analysis is becoming one of the most important fundamental works in financial content analysis.

We also observed a methodological trend in the literature. The body of sentiment analysis studies which use advanced statistical learning theory based machine learning methods is growing rapidly. This phenomenon is well exemplified by several recent eye-catching works (Li 2010, Huang and Li 2011, Antweiler and Frank 2005, Balakrishnan et al. 2010, Huang et al. 2010, Oh and Sheng 2011). A number of supervised machine learning methods were proposed and tested in these studies including Naïve Bayes, Support Vector Machines (SVM), Logistic Regression, K-Nearest Neighbors, etc. Since sentiment analysis is usually formalized as a text classification problem, it is not surprising that Naïve Bayes, one of the most widely used text classification methods, is the most prevalent one in these studies. During the data collection and processing phase of these studies, a common characteristic is that a great time and labor costs are required in data labeling simply because only labeled data can be used in training supervised learning models. Li (2010) hired 15 top BBA/MBA students to manually classify 30,000 sentences to support the model training. We can imagine that more hand-crafted efforts would be needed if larger scale experiments are to be conducted.

On one hand, it's encouraging to see that many interesting and insightful results have been reported in the literature as a result of using machine learning methods. On the other hand, it's also worthwhile to notice that several limitations of the current studies signify the preliminary stage of the application of machine learning methods in financial content analysis. First of all, though the body of research uses machine learning methods is growing, there is a scarcity of extensive theoretic analysis in the studies to indicate when and why a particular method would work well. Without such insight one may suffer from the seemingly misleading experiment results especially when the results are not as good as expected. Thus, when the unit of analysis is large and the structure within the data is complex (which is true in most of the real life applications), it's more likely to end up with non-optimal or even unexpectedly incorrect results. Secondly, there is few paper in the literature provides practical suggestions in properly applying machine learning methods. We argue that these two limitations tend not to encourage other business researchers

to build their work on the existing research and work towards a cumulative research tradition in this field (Dubin 1978). Theory development becomes difficult, if not impossible under this circumstance (Benbasat and Zmud 1999). Thirdly, current studies only focus on the supervised learning approach to text classification, the advances in using unlabeled data, namely the semi-supervised learning approach are largely overlooked. In semi-supervised learning, one enjoys the potential to embrace unlabeled data which is ubiquitous and easy to obtain in most of the cases, to improve the classification performance (Nigamy et al. 1998, Cozman et al. 2003). While this technique is not entirely new in machine learning, the application in financial content analysis is novel and worth exploring because it has the potential to largely reduce hand-crafted data labeling efforts.

Motivated by the above discussion, this paper is the attempt to address these limitations. The limitations of the extant literature well summarized our view on the current research gaps. Considering machine learning methods will likely to be widely adopted in this area and sentiment analysis is also becoming increasingly important, we believe the work towards filling the identified research gaps would be significant for both researchers and practitioners.

Methodology

Theoretical Foundation, the Bias-Variance Trade-off

Statistical learning theory is arguably the most important and systematic theoretical framework underlying the contemporary machine learning methods. The successful development and application of many machine learning algorithms in various fields such as computer vision, speech recognition and text classification could be attributed to the advances of such theory (Vapnik 1998). As most of the recent literature in financial content analysis using machine learning essentially formalizes the problems into a text classification problem, we believe there is a call for analyzing the text classification algorithm and the results under the guidance of both statistical learning theory and the financial domain knowledge. Among a number of aspects of the theory, the view on bias and variance, the two fundamental competing constructs in statistical learning, can be used in designing and evaluating learning algorithms. We will apply such view later in the consequent experiments. A systematic introduction on the theory can be found in (Hastie et al. 2009). Figure 1 shows the trade-off between bias and variance in a pictorial view.

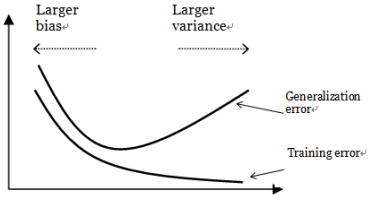

Figure 1. The Bias-variance Trade-off

Two Types of Naïve Bayes

As we mentioned earlier, naïve Bayes is one of the most widely used machine learning methods in the extant financial content analysis literature and in text mining literature in general. Machine learning literature shows that there are two distinct but interconnected variations of naïve Bayes, i.e., multinomial model and multivariate Bernoulli model (McCallum and Nigam 1998). It is subtle to notice this fact because the two possess different advantages and drawbacks for different scenarios. While incorrect application of the methods would cause considerable performance degeneration, it's either not clear or not explicitly mentioned in most of the financial content analysis literature that which variation of naïve Bayes was to be used. Since it's not entirely clear, from the statistical learning theory's view, which variation would outperform the other, we argue that it's very important to distinguish the two and test them with the real financial data.

Utilizing Unlabeled Data by Expectation Maximization Algorithm

We mentioned previously that current studies in financial content analysis only focus on using labeled data, namely the manually labeled sentences from the financial reports to train the text classification models. However, the heavy workload required in the data labeling would not only inevitably impose a restriction in scaling up the experiment, it also overlooks the opportunity to utilize the unlabeled data to further improve the classification performance. To address this, we need to switch to the semi-supervised learning paradigm. Expectation maximization (EM) algorithm is one of the most important algorithms in semi-supervised learning. We briefly describe how to use EM algorithm to enhance naïve Bayes algorithm to realize semi-supervised learning.

There are two steps in EM algorithm. The first step (the E-step) is to transfer the log-likelihood function to a more tractable equation and find a tight lower bound for it. The second step (the M-step) is to optimize the tractable equation by Maximum Likelihood Estimation (MLE) with respect to the parameters. Since EM is iterative by nature, we need to repeat the E-step and the M-step iteratively until the algorithm converges. It can be shown that EM is guaranteed to converge to a local optimum. More technical details of the algorithm and semi-supervised learning can be found in (Nigamy et al. 1998) and (Hastie et al. 2009).

Data Collection and Labeling

We collected data from two companies, Microsoft and Coach, for the following reasons. First of all, they are all leading companies in their own industries and their 10-K documents were very well prepared. This is important because we believe the result of the experiments with such data would generalize well to the data of other leading companies. Secondly, for the purpose of simulating real life applications we deliberately chose two companies from two very different industries. Because it's impossible to collect data from all the companies and label them all, it's very common for researchers to use data from one industry to predict the one from another industry. The classification difficulty in such setup is significant simply because both the writing style and the distribution of words are very different for companies from different industries. However, we created such scenario on purpose to explicitly examine the generalizability of the algorithm, namely to approach the performance of the algorithm in real applications. In addition, we also kept the amount of labeled data relatively small which is consistent with the situation in many real applications since high quality labeled data is always costly to obtain, especially when the labeling work requires specialized expertise.

In order to carry out both supervised learning and semi-supervised learning, our collected data can be divided into two portions, the labeled portion and the unlabeled portion. We first downloaded the 10-K documents for both Microsoft and Coach from the year of 2009 to 2012. We then extracted the MD&A section and manually labeled each sentence in this section into three different categories (i.e., positive, negative and neutral tone) for the year of 2011 and 2012. We regard the rest as the unlabeled data while the labeled portion can also be used as unlabeled data if we ignore the labels.

Table 1. The Statistics of the Final Data

	Microsoft (# of positive/negative/neutral sentences)	Coach (# of positive/negative/neutral sentences)
Year 2012	84/74/226 (22%/19%/59%)	108/26/194 (33%/8%/59%)
Year 2011	94/66/222 (25%/17%/58%)	106/30/190 (33%/9%/58%)
Year 2010	388 unlabeled sentences	301 unlabeled sentences
Year 2009	436 unlabeled sentences	317 unlabeled sentences

The output of the data labeling was verified and confirmed by a Certified Public Accountant (CPA) to ensure the labeling quality. Table 1 summarizes the statistics of the final data.

Experiments and Analysis

Comparing Naïve Bayes Variations

In the first experiment, we used the data from one company to train the model by which classifies the data collected from the same company. This scenario is very useful if one would like to study an individual company. We compared two naïve Bayes variations and the results are shown in the following figure.

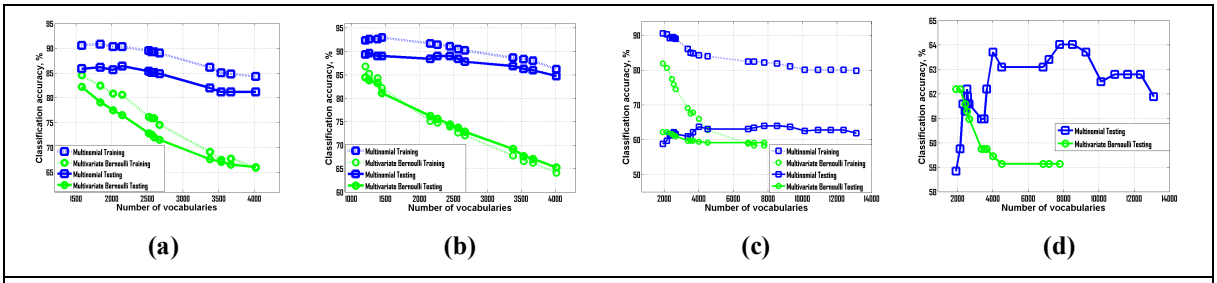

Figure 2. Comparison Between Two Naïve Bayes Variations
(a) To use Microsoft 2011 data to classify Microsoft 2012 data; (b) To use Coach 2011 data to classify Coach 2012 data; (c) To use Coach 2011 data to classify Microsoft 2012 data; (d) To use Microsoft 2011 data to classify Coach 2012 data

In figure 2 (a) and (b), the dotted lines represent the training accuracy and the solid lines represent the test accuracy. We used the following ways to explore the dynamics between bias and variance to seek a comprehensive overview of the performance. Since the amount of the labeled data is considered small, thus variance will be relatively big in this setting. One way to infuse more bias into the model is to increase the number of vocabularies in the dictionary. The reason why bias increases is that the enlarged dictionary contains more words which don't exist in the training set and it makes the model harder to fit the training data. As we indicated in figure 1, training accuracy is one indicator of the bias-variance trade-off, that's why we consider both training accuracy and test accuracy in the analysis.

It's noted from figure 2 (a) and (b) that the Microsoft case and the Coach case both exhibited the same trend in terms of the training accuracy, namely it decreased with the increase of the dictionary size. This makes sense because bigger dictionary provides larger bias which makes the data harder to fit. If we look at figure 1, we may say the algorithm is behaving towards the left (from high variance to high bias). Then we examined its corresponding effect on the test accuracy. According to the theory, the expected results should resemble the behavior of the generalization error curve in figure 1, namely the test accuracy should enjoy an increase first but eventually decrease due to the overwhelming bias. However, it's quite noticeable that the behavior of the test accuracy was very similar to the training accuracy and it didn't increase. Although it seems a bit surprising at the first glance, it's actually very reasonable. The main reason is the training data and the test data are very similar in this experiment, therefore the results on the test data can't be treated as a typical measure of the generalization error. Such similarity also makes the trend in the training accuracy resembles the test accuracy.

Though we didn't formally prove whether the multinomial naïve Bayes is guaranteed to outperform the multivariate Bernoulli model in this setting, the results suggested that the performance of the multivariate Bernoulli model is significantly decreased by the bias infused to the model while the performance of the multinomial model is very competitive. This implies that with the same amount of bias, the multinomial model seems to fit the data better. This phenomenon is consistent with the empirical evidence in the literature which indicates the multinomial model has stronger ability to fit the data. On the contrary, multivariate Bernoulli model is more sensitive to the bias (McCallum and Nigam 1998). The phenomenon is obvious in the Coach case in which the test accuracy became even higher than the training accuracy. This is an indicator suggesting the model is under-fitting the data.

We then used the model trained by one company's data to classify the data from another company. This scenario is of great interest because we usually can't collect data from all the companies we would like to investigate. We argue that the behavior of the test accuracy in this setting is a reasonable approximation to the generalization error because we are trying to generalize the model to the data it never sees before.

The results are showed in figure 2 (c) and (d). In addition to the similar findings on the training accuracy, we also observed two interesting results on the test accuracy. Firstly, the test accuracy of multivariate Bernoulli model is better than the multinomial model when the dictionary size is small. When we increased the dictionary size, the multinomial model dominated the test accuracy again. Secondly, we observed, for the multinomial model, that the test accuracy generated results resemble the curve of generalization error in figure 1. We better visualize this in figure 2 (d) by removing the curves of training accuracy.

The first result can be explained as follows. Initially, variance was high in the model due to the small dictionary size. However, the ability of better fitting the data made the multinomial model possess even greater variance. This led to the issue of over-fitting which caused low test accuracy. For multivariate Bernoulli model, the inability to well fit the data actually benefited the test accuracy because it provided a better trade-off in bias and variance against such dictionary size. However, with the increase of the dictionary size, the increased bias improved the bias-variance trade-off in the multinomial model, but hurt the multivariate Bernoulli model by introducing too much bias. Eventually, performance of both models will degenerate because too much bias hurts even for the multinomial model. The result is consistent with the theory illustrated in figure 1.

Utilizing Unlabeled Data

Though we gained some insights on the behaviors of different naïve Bayes models from the previous experiments, we observed that the performance in test accuracy would inevitably degenerate if we use the model to classify the data from a very different company. Now we seek to use both labeled and unlabeled data to improve the test accuracy in this scenario. This experiment is also interesting because unlabeled data is cheap and ubiquitous. The previous experiments showed that the multinomial variation generates better result with larger bias. Since adding unlabeled data would introduce dramatic change in the bias-variance dynamics, we carried out the experiment with the multinomial model because it's more robust to such changes from the previous experiments. In this experiment, the label led data was from Coach for the year of 2011 and 2012. We then trained the model with the additional unlabeled data from Microsoft of year 2009, 2010, 2011 and 2012 and tested with the data from Microsoft of year 2011 and 2012. This ended up with 1591 unlabeled samples. We then added more unlabeled from Coach of year 2011 and 2012 by ignoring the labels and trained the model again. This ended up with 2245 unlabeled samples. We reported the test accuracy in the following figure.

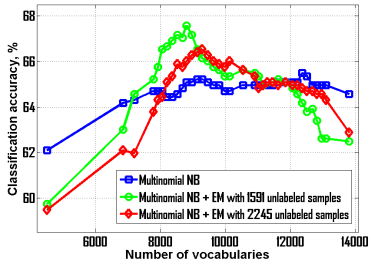

Figure 3. To use Coach 2011, 2012 labeled data plus unlabeled data to classify Microsoft 2011, 2012 data

We have a few interesting findings in this experiment. First of all, with the increase of bias introduced by the enlarging dictionary size, all of the three curves demonstrate the performance increase at the beginning and eventually suffer from the performance degeneration. This is consistent with the theory showed in figure 1. Secondly, once the bias-variance trade-off was well balanced, the utilization of unlabeled data significantly improved the classification performance. Thirdly, if the unlabeled data is added, it seems the model is more sensitive to the bias-variance trade-off, namely if the trade-off is not properly addressed, unlabeled data would cause the performance to degenerate faster. Last but not least, more unlabeled data seems to postpone the optimal performance interval. This could be attributed to the different bias-variance dynamics introduced by the difference in unlabeled sample sizes. These findings can be treated as empirical evidences for an important question, namely how and to what extend unlabeled data helps in sentiment analysis.

Discussions

Practical Suggestions for Sentiment Analysis Using Naïve Bayes

In the light of the experiments and analysis we presented, we may conclude some practical suggestions for both researchers and practitioners in applying naïve Bayes algorithm in financial content analysis. 1). It's very important to always consider the bias-variance trade-off. Unexpected results are likely to occur if such trade-off is ignored. 2). When the training data and the test data are relatively similar (e.g., a company's historic data is used to classify its recent data), multinomial naïve Bayes is preferred. Meanwhile, smaller dictionary size is likely to be superior to the larger ones in this particular setting. 3). If you would like to generalize the model to completely novel data, the multivariate Bernoulli naïve Bayes tend to work better with smaller dictionary size. If you have to use big dictionary size, you should choose multinomial naïve Bayes. 4). It's useful to keep in mind that performance improvement is not guaranteed by adding unlabeled data. Balancing the bias-variance trade-off is the way to achieve better performance.

Contributions and Limitations

The contributions of this paper come from what we've done for the identified research gaps. Firstly, we drew on statistical learning theory to provide theoretic analysis on when and why naïve Bayes, either with or without unlabeled data, generates preferred performance. We also provided practical suggestions in applying naïve Bayes methods in financial content analysis. We argue that this not only contributes to the IS literature by accumulating fundamental knowledge, it also encourages the cumulative manner in the literature which is important to theory building. We believe this is especially necessary for the inter-disciplinary fields such as IS. Secondly, we found that when the amount of labeled data is limited, the utilization of unlabeled data improves the classification performance if the bias-variance trade-off is well balanced. This result has the potential to largely reduce the workload of the tedious data labeling and may facilitate the scaling up of the experiment in financial content analysis.

There are a few limitations in this study. Firstly, we only collected data from two companies. Though we seek the generalizability of our results by collecting data from two very different leading companies, there is still work to be done to confirm the external validity of the results. Secondly, we found that if the bias-variance trade-off is well balanced, unlabeled data would improve the classification performance. However, we didn't mention how to automatically achieve this balance as well as to what extend unlabeled data would help. This may be achieved by studying a number of variations of feature selection methods. We observed several critical discussions on whether and when unlabeled data would help in the machine learning literature (Nigamy et al. 1998, Cozman et al. 2003). It's not a trivial task to show if the issue could be addressed by better balancing the bias-variance trade-off. Since the absence of it wouldn't compromise the goal of this study, we would leave it to the future work.

Conclusions and Future work

In this paper, we identified several research gaps in applying machine learning methods to financial sentiment analysis. We emphasized the importance of statistical learning theory and argued that both theoretic analysis and practice guidance in applying machine learning methods would not only generate useful research output, it encourages the cumulative manner in the literature as well. In addition, though the unlabeled data should be utilized carefully, we showed it has the ability to improve sentiment analysis performance and the potential to alleviate the manual data labeling work.

Future work may include the following aspects. Additional studies with data from other companies should be carried out to validate the results of this study. It's also necessary and important to examine other text mining methods to see if the results of this paper would also be valid. Besides, it would be very interesting to see if unlabeled data would help in scaling up the experiments in this field. It's also an interesting direction to examine whether unlabeled would always help given the bias-variance trade-off is carefully balanced.

Acknowledgements

We thank the anonymous reviewers and the editors for providing us valuable comments and suggestions.

References

Antweiler, W. and M. Z. Frank. 2005. "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards," *The Journal of Finance* (59:3), pp. 1259–1294.

Balakrishnan, R., X. Y. Qiu, et al. 2010. "On the predictive ability of narrative disclosures in annual reports," *European Journal of Operational Research* (202:3), pp. 789–801.

Benbasat, I. and R. W. Zmud. 1999. "Empirical research in information systems: The practice of relevance," *MIS Quarterly* (23:1), pp. 3-16.

Bryan, S. H. 1997. "Incremental Information Content of Required Disclosures Contained in Management Discussion and Analysis," *The Accounting Review* (72:2), pp. 285-301.

Cozman, F. G., I. Cohen, et al. 2003. "Semi-Supervised Learning of Mixture Models," In *Proceedings of 20th International Conference on Machine Learning (ICML)*. Washington, DC.

Davis, A. K., J. M. Piger, et al. 2012. "Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language," *Contemporary Accounting Research* (29:3), pp. 845–868.

Dubin, R. 1978. *Theory Development*, New York, Free Press.

Hastie, T., R. Tibshirani, et al. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, Springer.

Huang, A., A. Zang, et al. 2010. "Informativeness of Text in Analyst Reports: A Naïve Bayes Machine Learning Approach," HKUST work in progress.

Huang, K.-W. and Z. Li. 2011. "A Multilabel Text Classification Algorithm for Labeling Risk Factors in SEC Form 10-K," *ACM Transactions on MIS* (2:3), Article 18.

Humpherys, S. L., K. C. Moffitt, et al. 2011. "Identification of fraudulent financial statements using linguistic credibility analysis," *Decision Support Systems* (50:3), pp. 585–594.

Kothari, S. P., X. Li, et al. 2009. "The Effect of Disclosures by Management, Analysts, and Financial Press on the Equity Cost of Capital: A Study Using Content Analysis," *Accounting Review of Accounting Studies* (84:5), pp. 1639–1670.

Kravet, T. and V. Muslu (Forthcoming). "Textual Risk Disclosures and Investors' Risk Perceptions," Review of Accounting Studies.

Li, F. 2010. "The Information Content of Forward-Looking Statements in Corporate Filings—A Naive Bayesian Machine Learning Approach," *Journal of Accounting Research* (48:5), pp. 1049-1102.

McCallum, A. and K. Nigam. 1998. "A Comparison of Event Models for Naive Bayes Text Classication," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-98)*. Madison, Wisconsin.

Nigamy, K., A. McCallum, et al. 1998. "Learning to Classify Text from Labeled and Unlabeled Documents," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-98)*. Madison, Wisconsin.

Oh, C. and O. Sheng. 2011. "Investigating Predictive Power of Stock Micro Blog Sentiment in. Forecasting Future Stock Price Directional Movement," in *Proceedings of the* International Conference on Information Systems (ICIS 2011). Shanghai, China.

Tetlock, P. C., M. Saar-Tsechansky, et al. 2008. "More Than Words: Quantifying Language to Measure Firms' Fundamentals," *The Journal of Finance* (63:3), pp. 1437–1467.

Vapnik, V. 1998. *Statistical Learning Theory*, Wiley-Interscience.